Update Estimation and Scheduling for Over-the-Air Federated Learning with Energy Harvesting Devices

Furkan Bagci[1], Busra Tegin[1], Mohammad Kazemi[2] and Tolga M. Duman[1]
[1]*Dept. of Electrical and Electronics Engineering, Bilkent University*, Ankara, Turkey
[2]*Dept. of Electrical and Electronic Engineering, Imperial College London*, London, UK
{bagci, btegin, duman}@ee.bilkent.edu.tr, mohammad.kazemi@imperial.ac.uk

***Abstract*—We study over-the-air (OTA) federated learning (FL) for energy harvesting devices with heterogeneous data distribution over wireless fading multiple access channel (MAC). To address the impact of low energy arrivals and data heterogeneity on global learning, we propose user scheduling strategies. Specifically, we develop two approaches: 1) entropy-based scheduling for known data distributions and 2) least-squares-based user representation estimation for scheduling with unknown data distributions at the parameter server. Both methods aim to select diverse users, mitigating bias and enhancing convergence. Numerical and analytical results demonstrate improved learning performance by reducing redundancy and conserving energy.**

***Index Terms*—Federated learning, over-the-air, fading MAC, energy harvesting, device scheduling, least-squares estimation.**

I. Introduction

With the growing number of devices collecting high-quality data, vast amounts of practical information are available for machine learning (ML) applications. Traditional ML approaches require centralized data sharing, demanding significant resources and raising privacy concerns. To address these issues, *Federated Learning* (FL) enables collaborative model training without sharing local data, ensuring privacy, low latency, and improved learning quality [1], [2]. However, FL relies on iterative transmission of local updates from mobile users (MUs) to the parameter server (PS), making communication bandwidth a key bottleneck [3]. Over-the-air (OTA) computation addresses this issue by leveraging wireless multiple access channel (MAC) to aggregate updates directly during transmission [3], [4]. Multiple receive antennas with combining techniques in OTA FL can mitigate channel effects and alleviate fading, enabling convergence even with blind transmitters lacking channel state information (CSI) [5].

The deployment of FL faces many challenges, such as hardware impairments, data distribution issues, and limited energy for MUs. To address energy constraints, energy harvesting (EH) devices are widely used to obtain energy from the environment, with studies exploring the channel capacity with various battery sizes [6]. For FL with OTA EH devices, studies have examined joint user selection and receive beamforming [7], weighted averaging to prevent bias [8], transceiver optimization [9], and Markov decision processes for joint scheduling and power management [10].

Studies on FL with non-independent and identically distributed (non-i.i.d.) data show that heterogeneous local datasets significantly impact model accuracy and convergence [11], [12]. In [13], a deep reinforcement learning-based user selection framework is utilized to mitigate bias by exploiting the relationships between user updates and data distributions. In [14], clustered sampling is used to schedule users based on cosine similarity, improving representation. In [15], federated averaging with diverse user selection is employed to approximate full participation gradients, reducing redundancy from similar data. These studies assume user updates reflect data distributions, relying on separate transmission of user updates. In contrast, our OTA FL setup uses noisy aggregated updates transmitted simultaneously by all scheduled users, eliminating the need for separate-transmission resource allocation.

In this work, we explore diverse user selection for FL with energy harvesting devices and heterogeneous data distribution by scheduling active users based on their data distribution using over-the-air transmission. Firstly, we select a subset of users that achieves a uniform data distribution using exact data distribution information to ensure the aggregate of selected users' updates approximates that of all users. Next, since knowing data distribution raises privacy concerns, we estimate user updates from aggregated signals on the PS via least-squares estimation (LSE), enabling user selection without revealing data distribution information. Unlike existing works, we use OTA aggregated signals instead of individual user updates for transmission efficiency along with participation information to estimate local model update representations. These representative updates are utilized to group users into clusters, improving learning performance and identifying redundant information in the system.

The paper is organized as follows. Section II introduces the FL setup and EH devices. Section III analyzes OTA FL for EH devices with a unit-sized battery and provides convergence analysis. In Section IV, we propose update estimation and user scheduling policies, with numerical results presented in Section V. We conclude the paper in Section VI.

Furkan Bagci's work is supported by Türk Telekomünikasyon A.S. within the framework of the 5G and Beyond Joint Graduate Support Programme coordinated by the Information and Communication Technologies Authority.

The authors acknowledge support from TUBITAK through CHIST-ERA project SONATA (CHIST-ERA-20-SICT-004, funded by TUBITAK, Turkey Grant 221N366).

The work of Mohammad Kazemi was supported by UKRI under the U.K. government's Horizon Europe Funding Guarantee under Grant 101103430.

Notations: For vectors $\mathbf{x}$ and $\mathbf{y}$ of the same dimension, $\mathbf{x}\circ\mathbf{y}$ denotes element-wise product. We define $[i]=\{1,\dots,i\}$.

II. System Model

We consider an FL system with heterogeneous data distribution for M energy harvesting devices. We assume that the mobile users do not have CSI and transmit their updates to the PS through a fading MAC using over-the-air transmission. The PS employs K receive antennas to align the received signals in the absence of CSI at the transmitters (CSIT) using aggregated channel information.

In FL, the primary objective is to minimize a global loss function, denoted as $F(\boldsymbol{\theta})$, collaboratively across M devices, where $\boldsymbol{\theta}\in\mathbb{R}^{2N}$ represents the parameters of the global model to be optimized. The global loss function is defined as

$$F(\boldsymbol{\theta})=\frac{1}{B}\sum_{m=1}^{M}\frac{|\mathcal{B}_m|}{B}F_m(\boldsymbol{\theta}), \tag{1}$$

where $F_m(\boldsymbol{\theta})$ represents the average empirical local loss for the m-th user with model parameters $\boldsymbol{\theta}$. For m-th user with local dataset $\mathcal{B}_m$, for $m\in[M]$, and $B\triangleq\sum_{m=1}^{M}|\mathcal{B}_m|$, we have

$$F_m\left(\boldsymbol{\theta}\right)=\frac{1}{|\mathcal{B}_m|}\sum_{\boldsymbol{u}\in\mathcal{B}_m}f\left(\boldsymbol{\theta},\boldsymbol{u}\right), \tag{2}$$

with $f(\boldsymbol{\theta},u)$ as the empirical loss function corresponding to the u-th data sample in the local dataset B_m.

In FL with EH devices, unlike traditional FL, the limited energy availability can cause some users to lack sufficient energy to perform local computations and transmissions. Consequently, contributions will only come from the MUs that have enough energy and are selected based on the scheduling policy. At each global iteration t, the PS broadcasts the latest global model, $\boldsymbol{\theta}(t)$. In response, the selected MUs perform τ iterations of local stochastic gradient descent (SGD) to minimize their individual local loss functions, $F_m(\boldsymbol{\theta})$, for $m\in\mathcal{S}(t)$, where $\mathcal{S}(t)$ is the set of scheduled users. Subsequently, the model updates obtained by the mobile users are transmitted back to the PS to contribute to the global learning process.

To compute the local model updates, the m-th user at the i-th local and t-th global iteration performs the local iterations of SGD with the following update rule:

$$\boldsymbol{\theta}_m^{i+1}(t)=\boldsymbol{\theta}_m^{i}(t)-\eta_m^i(t)\nabla F_m(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)), \tag{3}$$

where $i\in[\tau]$, $\eta_m^i(t)$ is the learning rate and $\nabla F_m\left(\boldsymbol{\theta}_m^i(t),\xi_m^i(t)\right)$ represents the stochastic gradient estimate for the $\boldsymbol{\theta}_m^i(t)$ and the local mini-batch sample ξ_m^i randomly chosen from the local dataset $\mathcal{B}_m$.

After completing the local SGD steps, the m-th user computes the model update, which is aimed to be shared with the PS as

$$\Delta\boldsymbol{\theta}_m(t)=\boldsymbol{\theta}_m^{\tau}(t)-\boldsymbol{\theta}_m^{1}(t). \tag{4}$$

Using over-the-air transmission over a fading MAC, the received signal at the k-th antenna of the PS at iteration t is given as

$$\boldsymbol{y}_{PS,k}(t)=\sum_{m\in\mathcal{S}(t)}\boldsymbol{h}_{m,k}(t)\circ\boldsymbol{x}_m(t)+\boldsymbol{z}_{PS,k}(t), \tag{5}$$

where $\boldsymbol{x}_m(t)$ is the signal transmitted by the m-th user, and $\boldsymbol{h}_{m,k}(t)$ is the independent and identically distributed (i.i.d.) channel gains from the m-th user to the k-th antenna with entries $h_{m,k}^n(t)\sim\mathcal{CN}(0,\sigma_h^2)$. Similarly, $z_{PS,k}^n(t)$ denotes the n-th entry of the channel noise, $\boldsymbol{z}_{PS,k}(t)$, which is i.i.d. circularly symmetric additive white Gaussian noise (AWGN) and is distributed according to $\mathcal{CN}(0,\sigma_z^2)$.

The PS uses the received signals from K antennas to update the global model as

$$\boldsymbol{\theta}_{PS}(t+1)=\boldsymbol{\theta}_{PS}(t)+\Delta\hat{\boldsymbol{\theta}}_{PS}(t), \tag{6}$$

where $\boldsymbol{\theta}_{\text{PS}}(t)$ represents the global model vector at the server at global iteration t and $\Delta\hat{\boldsymbol{\theta}}_{PS}(t)$ is the noisy estimate of the average of the local updates. Note that if there were no noise or fading, the average of the local updates would be

$$\Delta\boldsymbol{\theta}_{PS}(t)=\frac{1}{|\mathcal{S}(t)|}\sum_{m\in\mathcal{S}(t)}\Delta\theta_m(t). \tag{7}$$

A. EH Devices

We study FL with OTA for EH devices, each with a unit-sized battery. At each global iteration, devices harvest energy with varying success, storing it for future use following the harvest-store-use approach [16]. Surplus energy is lost if the battery is full during energy arrivals. Local SGD and update transmissions consume one energy unit global per iteration, highlighting stochastic energy availability as a key constraint.

We consider a Bernoulli energy arrival process where, at each global iteration t, the m-th user receives unit energy with probability $p_e^m(t)$. Active users consume available battery energy, while inactive users store it for future use. Energy arrivals are shared with the PS after each iteration, as in [7], [9].

III. OTA FL with EH Devices with Unit Battery

In this section, we consider FL with OTA aggregation, where only active users contribute to the iterations due to limited energy arrivals and provide a brief convergence analysis.

The model updates of the scheduled users, $\Delta\boldsymbol{\theta}_m^{cx}(t)\in\mathbb{C}^N$, $m\in\mathcal{S}(t)$, are transmitted as complex signals at iteration t, represented as

$$\Delta\boldsymbol{\theta}_m^{re}(t)\triangleq\left[\Delta\theta_m^1(t),\Delta\theta_m^2(t),\dots,\Delta\theta_m^N(t)\right]^T, \tag{8a}$$

$$\Delta\boldsymbol{\theta}_m^{im}(t)\triangleq\left[\Delta\theta_m^{N+1}(t),\Delta\theta_m^{N+2}(t),\dots,\Delta\theta_m^{2N}(t)\right]^T, \tag{8b}$$

$$\Delta\boldsymbol{\theta}_m^{cx}(t)\triangleq\Delta\boldsymbol{\theta}_m^{re}(t)+j\Delta\boldsymbol{\theta}_m^{im}(t). \tag{8c}$$

Using the channel output at each antenna as defined in (5), the PS combines signals from K antennas using the sum of the channel gains from the scheduled users to each antenna as

$$\boldsymbol{y}_{PS}(t)=\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{m\in\mathcal{S}(t)}\boldsymbol{h}_{m,k}(t)\right)^*\circ\boldsymbol{y}_{PS,k}(t). \tag{9}$$

where the received signals $\boldsymbol{y}_{PS,k}(t)$'s are given in (5) for the transmitted signal $\Delta\boldsymbol{\theta}_m^{cx}(t)$.

The n-th symbol of (9) can be partition into three signals

$$y_{PS}^n(t) = \underbrace{\sum_{m\in\mathcal{S}(t)} \left(\frac{1}{K}\sum_{k=1}^{K}|h_{m,k}^n(t)|^2\right)\Delta\theta_m^{n,cx}(t)}_{y_{PS}^{n,sig}(t)\text{(signal term)}}$$
$$+\underbrace{\frac{1}{K}\sum_{m\in\mathcal{S}(t)}\sum_{\substack{m'\in\mathcal{S}(t)\\ m'\neq m}}\sum_{k=1}^{K}(h_{m,k}^n(t))^* h_{m',k}^n(t)\Delta\theta_{m'}^{n,cx}(t)}_{y_{PS}^{n,int}(t)\text{(interference term)}}$$
$$+\underbrace{\frac{1}{K}\sum_{m\in\mathcal{S}(t)}\sum_{k=1}^{K}(h_{m,k}^n(t))^* z_{PS,k}^n(t)}_{y_{PS}^{n,noise}(t)\text{(noise term)}}. \tag{10}$$

As shown in [5], the variance of the interference coefficient $\Delta\theta_{m'}^{n,cx}(t)$ decreases with the number of antennas K. Hence, a sufficient number of antennas allows for accurate estimation and recovery of noisy aggregated updates as

$$\Delta\hat{\boldsymbol{\theta}}_{PS}^n(t) = \frac{1}{|\mathcal{S}(t)|\,\sigma_h^2}\mathrm{Re}\{y_{PS}^n(t)\}, \tag{11a}$$
$$\Delta\hat{\boldsymbol{\theta}}_{PS}^{n+N}(t) = \frac{1}{|\mathcal{S}(t)|\,\sigma_h^2}\mathrm{Im}\{y_{PS}^n(t)\}, \tag{11b}$$

to update the global model, as in (6).

A. Convergence Analysis

In this section, we provide convergence analysis for the OTA FL with EH devices and no CSIT by upper-bounding the gap between our model estimate and the optimal model.

The optimal solution minimizing (1) is $\boldsymbol{\theta}^* \triangleq \arg\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, with optimal loss $F^* = F(\boldsymbol{\theta}^*)$. For user $m \in [M]$, the optimal local model is $\boldsymbol{\theta}_m^* \triangleq \arg\min_{\boldsymbol{\theta}_m} F_m(\boldsymbol{\theta}_m)$, with corresponding loss $F_m^* = F(\boldsymbol{\theta}_m^*)$.

1) Preliminaries: The amount of bias and heterogeneity across devices is represented by the following non-negative parameter

$$\Gamma = F^* - \sum_{m=1}^{M}\frac{B_m}{B}F_m^*, \tag{12}$$

A higher Γ indicates significant non-i.i.d. data distribution, while $\Gamma \to 0$ reflects near i.i.d..

We consider the same learning rate across users and local iterations, $\eta_m^i(t) = \eta(t)$, but allow it to vary between global iterations. The local model update at the m-th user for global iteration t and local iteration $i \in [\tau]$ is given as

$$\boldsymbol{\theta}_m^{i+1}(t) - \boldsymbol{\theta}_m^1(t) = -\eta(t)\sum_{l=1}^{i}\nabla F_m\left(\boldsymbol{\theta}_m^l(t), \xi_m^l(t)\right). \tag{13}$$

To perform the convergence analysis, similar to the existing studies [5], [17], [18], we assume that the loss functions $F_1, \ldots, F_M$ are all L-smooth and μ-strongly convex. Also, it is assumed that the expected squared ℓ_2-norm of the stochastic gradients is bounded; that is, for all $i \in [\tau]$, $m \in [M]$, and t, we have $\mathbb{E}_\xi\left[\left\|\nabla F_m\left(\boldsymbol{\theta}_m^i(t), \xi_m^i(t)\right)\right\|_2^2\right] \leq G^2$.

2) Convergence Rate: For the convergence analysis of OTA FL with EH devices and data heterogeneity, we begin by analyzing the convergence by comparing the optimal model with our system model, where only a subset of users participate in each iteration. Using these findings, we guide user scheduling decisions to minimize the discrepancy in this bound. Our main result is as follows.

Theorem 1. *For $0 < \eta(t) \leq \min\left\{1, \frac{1}{\mu\tau}\right\}, \forall t$. We have*

$$\mathbb{E}\left[\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\|_2^2\right] \leq \left(\prod_{i=0}^{t-1}A(i)\right)\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\|_2^2 + \sum_{j=0}^{t-1}B(j)\prod_{i=j+1}^{t-1}A(i), \tag{14}$$

with

$$A(i) \triangleq 1 - \mu\eta(i)\ \left(\tau - \eta(i)(\tau-1)\right), \tag{15}$$
$$\begin{aligned} B(i) \triangleq\ & \frac{\eta^2(i)\tau^2G^2}{K} + \frac{\sigma_z^2 N}{\alpha_i^2 K|\mathcal{S}(i)|\sigma_h^2} \\ & + (1+\mu(1-\eta(i)))\,\eta^2(i)G^2\frac{\tau(\tau-1)(2\tau-1)}{6} \\ & + \eta^2(i)(\tau^2+\tau-1)G^2 + 2\eta(i)(\tau-1)\Gamma \\ & + \left(\eta^2(i)\tau(\tau-1)LG + \eta(i)\tau\epsilon\right)^2 \\ & + \left(\eta^2(i)\tau(\tau-1)LG + \eta(i)\tau\epsilon\right)c, \end{aligned} \tag{16}$$

for some constant $c \geq 0$.

Proof. See Appendix A. □

We note that A(i) represents the decay rate of the distance from the initial starting point to the optimal solution. In $B(i)$, the first two terms represent the transmission error due to the wireless fading MAC with blind transmitters, and the third and fourth terms are related to federated averaging. Additionally, we emphasize that the last two terms represent the error caused by partial user participation similar to [15], with ϵ in (16) being the gradient approximation error and defined as follows

$$\epsilon \triangleq \left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F_m(\theta_m(t)) - \frac{1}{|S(t)|}\sum_{m\in\mathcal{S}(t)}\nabla F_m(\theta_m(t))\right\|_2. \tag{17}$$

Via user scheduling, our aim is to select the subset of users with the average local updates closest to the full participation case, thereby minimizing the error ϵ in (17). Hence, in the next section, we propose user selection and scheduling approaches that minimize the distance between the trained FL model and the optimal model, even in scenarios involving EH devices and highly non-i.i.d. data distributions.

IV. Update Estimation and User Scheduling

In this section, we demonstrate that the data distribution of users is critical in the scheduling procedure, especially for highly non-i.i.d data distributions, and can be used to minimize the error bound related to the partial participation characteristics of EH devices. First, we propose an entropy-based user scheduling policy with known data distribution. Then, we extend our discussion to the unknown data distribution case and show that the data distribution characteristics can be estimated via least-squares to be used in user scheduling.

A. Entropy-based User Scheduling with Known Data Distributions

Assuming all MUs reveal their data distribution to the PS beforehand, our method selects a subset of users that effectively represents all data labels in the network. Based on this, the PS characterizes the label distribution of each user $m \in [M]$ as $L_m = [l_{m,0}, l_{m,1}, \ldots, l_{m,N_c-1}]$, where N_c is the total number of classes, and l_{m,n_c} represents the portion of the m-th user's data corresponding to label n_c. At each iteration, the PS computes the label distribution for all available user subsets as a probability mass function and selects the one with the highest Shannon entropy, indicating the most balanced label distribution available. While this strategy is similar to that in [19], we extend our approach to a more practical setup that incorporates OTA transmission, wireless fading MAC, and blind transmitters, showing the effectiveness of entropy-based user selection for EH devices under practical constraints.

B. User Clustering and Scheduling with Unknown Data Distribution

We consider a more realistic case, where the PS lacks data distribution information, ensuring a level of user privacy.

For user scheduling, we rely on the relationship between the model updates from users and their underlying data distribution, similar to studies in [13]–[15]. Unlike these studies, our approach is constrained to using a noisy estimate of the sum of updates provided by all selected users at a given iteration. We demonstrate that a representation of the user updates can be estimated at the PS, allowing users to be grouped into clusters based on the similarities between their representations to minimize error from partial participation in (17). This approach selects suitable users while preventing redundant information transfer and conserving energy under the constraints of EH devices.

To achieve this, we use LSE to create a representation of the updates. Over T estimation iterations, the PS stores normalized global updates from (11) while all the active users participate without scheduling, termed the *estimation phase*. Note that we normalize user updates to unit norm to mitigate scale discrepancies. At the end of this estimation window, PS estimates representative updates based on stored global updates and participation information. We emphasize that the goal is to estimate a representation of user updates rather than recovering the individual updates themselves.

We define a matrix $\hat{\boldsymbol{\Theta}}_{PS}$, whose rows represent global model updates $\Delta\hat{\boldsymbol{\theta}}_{PS}(t)$ from (11). For the j-th iteration with $j \leq T$, the j-th row of this matrix can be expressed as $\hat{\boldsymbol{\Theta}}_{PS,j} = \boldsymbol{A}_j\boldsymbol{\Theta}_j + \boldsymbol{N}_j^{'}$, where $\boldsymbol{A}_j$ is a binary participation vector with $\boldsymbol{A}_j \in \{0,1\}^{1\times M}$, and $\boldsymbol{\Theta}_j \in \mathbb{R}^{M\times 2N}$, with each row representing the local model update for a specific user $m \in [M]$, denoted as $\Delta\boldsymbol{\theta}_{j,m}$. Additionally, $\boldsymbol{N}_j^{'} \in \mathbb{R}^{1\times 2N}$, whose d-th element is denoted by $N_{j,d}^{'}$ for $d \in [2N]$, represents the effective noise arising from MAC fading, AWGN, and PS combining errors. $\hat{\boldsymbol{\Theta}}_{PS,j}$ is expressed as:

$$\hat{\boldsymbol{\Theta}}_{PS,j} = \boldsymbol{A}_j \begin{bmatrix} \Delta\boldsymbol{\theta}_{j,1} \\ \vdots \\ \Delta\boldsymbol{\theta}_{j,M} \end{bmatrix} + \begin{bmatrix} N_{j,1}^{'} & \ldots & N_{j,2N}^{'} \end{bmatrix}. \tag{18}$$

We also define $\boldsymbol{\Theta}_{rep} \in \mathbb{R}^{M\times 2N}$ as a representation of local updates. Using this, $\hat{\boldsymbol{\Theta}}_{PS,j}$ can be written as:

$$\hat{\boldsymbol{\Theta}}_{PS,j} = \boldsymbol{A}_j(\boldsymbol{\Theta}_{rep} + \boldsymbol{\Theta}_{diff,j}) + \boldsymbol{N}_j^{'}, \tag{19}$$

where $\boldsymbol{\Theta}_{diff,j}$ is defined as the difference between $\boldsymbol{\Theta}_{rep} - \boldsymbol{\Theta}_j$. Combining (19) for $j \in \{1, \ldots, T\}$ and defining a total noise term $\boldsymbol{N}_j^* \triangleq \boldsymbol{A}_j\boldsymbol{\Theta}_{diff,j} + \boldsymbol{N}_j^{'}$, which represents the noise due to the channel, interference from the blind transmitters, and the difference between representative updates and the real updates, we obtain

$$\hat{\boldsymbol{\Theta}}_{PS} = \boldsymbol{A}\boldsymbol{\Theta}_{rep} + \boldsymbol{N}^*, \tag{20}$$

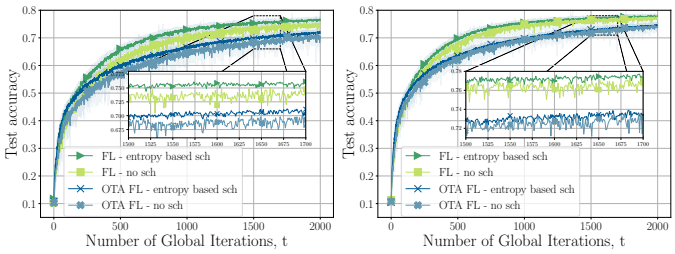
where $\hat{\boldsymbol{\Theta}}_{PS} = [\Delta\hat{\boldsymbol{\theta}}_{PS}(t-T+1); \cdots; \Delta\hat{\boldsymbol{\theta}}_{PS}(t)] \in \mathbb{R}^{T\times 2N}$ and ';' represents row-wise concatenation. Additionally, we have $\boldsymbol{A} \in \{0,1\}^{T\times M}$, $\boldsymbol{\Theta}_{rep} \in \mathbb{R}^{M\times 2N}$ and $\boldsymbol{N}^* \in \mathbb{R}^{T\times 2N}$.

By solving LSE for (20), we can get an estimate for the representative updates as $\hat{\boldsymbol{\Theta}}_{rep}$. Using this representation, the PS can infer the characteristics of the users' data distribution as a similarity between user representations, which can then be utilized in the user selection procedure. Notably, the PS infers similarity among user representations without accessing their data distribution, preserving user privacy.

Due to the limited and stochastic energy arrivals, some users might dominate the training and create a bias towards certain labels and certain users. By employing *cosine similarity*, users are clustered into groups to promote diverse user contributions by selecting the expected number of users based on energy distribution from each cluster to ensure unbiased training. This approach helps to reduce the bias due to the non-i.i.d data and provides fair performance among clients, as noted similarly in [13], [15].

V. Numerical Results

In this section, we evaluate the performance of our proposed user scheduling methods across multiple scenarios. We consider image classification tasks on the MNIST [20], FMNIST [21], and CIFAR-10 [22] datasets under non-i.i.d. data distributions. For MNIST and FMNIST, we use a single-layer neural network with 784 input and 10 output neurons ($2N = 7850$). For CIFAR-10, we use a convolutional neural network (CNN) ($2N = 797962$) as in [23]. Training is performed by SGD



(a) $\beta = 0.1$. (b) $\beta = 0.2$.

Fig. 1: The mean test accuracy of entropy-based scheduling for CIFAR-10 with $M = 100$, $|\mathcal{B}_m| = 500$ and $p_e^m(t) = 0.1$, $\forall m, t$.

with a learning rate of 0.05 and a scheduler, $\tau = 5$ and mini-batch size $|\xi_m(t)| = 100$ for MNIST and FMNIST, and $\tau = 3$ and $|\xi_m(t)| = 128$ for CIFAR-10.

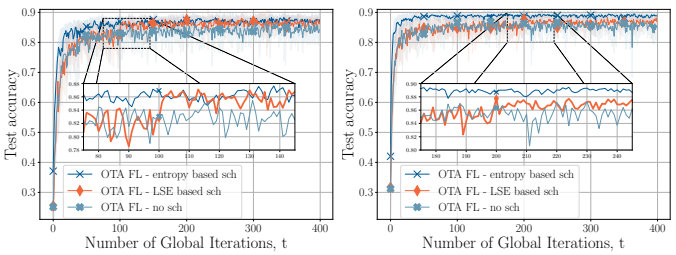
To simulate highly non-i.i.d. data, we consider two different distribution scenarios. In the first scenario, users receive data limited to a fixed number of labels, either 1 or 2 classes assigned per user. In the second one, we sample $\boldsymbol{p_m} \sim \text{Dir}_{N_c}(\beta)$ with $\boldsymbol{p_m} = [p_{m,0}, \cdots, p_{m,N_c-1}]$, and user m receives p_{m,n_c} portion of its data from class $n_c \in [N_c]$. β is a Dirichlet distribution parameter, where smaller values of β lead to more unbalanced partitions.

We evaluate the performance of our wireless OTA FL with highly non-i.i.d data against a baseline scenario without user scheduling, where users participate in the global learning process whenever they have energy. Throughout the simulations, users are connected to a PS through wireless fading MAC where channel gains from each user to each PS antenna are i.i.d., with parameters $K = 200$, $\sigma_h^2 = 1$, and $\sigma_z^2 = 0.1$.

In Fig. 1, we demonstrate the performance of entropy-based scheduling for CIFAR-10 dataset with $\beta \in \{0.1, 0.2\}$ for $M = 100$ users with $|\mathcal{B}_m| = 500$ and $p_e^m(t) = 0.1$ for $m \in [M]$. We observe that the gains from our scheme are more significant in scenarios with greater heterogeneity for both the error-free and OTA FL cases. As shown in the figure, as the distribution becomes more heterogeneous as in Fig 1a, (i.e., $\beta = 0.1$), the impact of entropy-based scheduling increases.

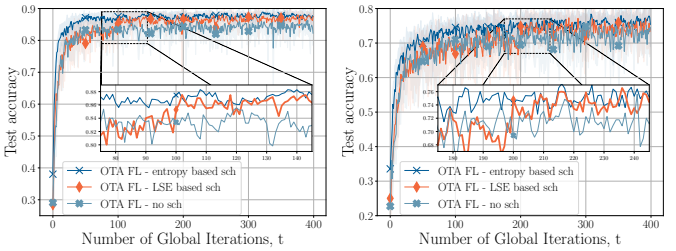
Fig. 2 shows the mean test accuracies for data distributions with 1 class and 2 classes per user on the MNIST dataset with $M = 40$, $|\mathcal{B}_m| = 1250$, $p_e^m(t) = 0.25$ for $m \in [M]$ and estimation phases of $T = 100$ and $T = 200$ iterations. In both cases, entropy-based scheduling results in higher and more stable accuracy levels. For the unknown data distribution, the PS estimates local user representations after T iterations and groups users into 10 clusters, scheduling one user per cluster for each iteration, similar to [13]. Similar to the previous case, our scheme achieves higher gains in more heterogeneous scenarios. Estimation and clustering become harder with less heterogeneity, yet scheduling still outperforms the no-scheduling baseline.

In Fig. 3, we illustrate the performance of our scheduling policies with 10 clusters for the MNIST dataset with $M = 20$ users, where $|\mathcal{B}_m| = 2500$, $p_e^m(t) = 0.5$ and FMNIST dataset with $M = 40$ users, where $|\mathcal{B}_m| = 1250$, $p_e^m(t) = 0.25$ for $m \in [M]$. For known data distributions, the entropy-based scheduling method outperforms the no-scheduling baseline



(a) 1 class per user and $T = 100$. (b) 2 class per user and $T = 200$.

Fig. 2: The mean test accuracy for MNIST with $M = 40$, $|\mathcal{B}_m| = 1250$ and $p_e^m(t) = 0.25$, $\forall m, t$.



(a) MNIST, $M = 20$, 1 class per user and $T = 100$. (b) FMNIST, $M = 40$, 1 class per user and $T = 200$.

Fig. 3: The mean test accuracy for MNIST and FMNIST.

approach. For unknown data distributions, the global model's performance improves after the estimation phase, approaching the entropy-based case.

The results above demonstrate that our entropy-based scheduling approach, leveraging users' data distributions, gives a performance boost for OTA FL with EH devices over a wireless MAC channel with blind transmitters, yielding higher and more stable accuracy levels. Moreover, the results demonstrated that in cases of unknown data distributions, user representations can be estimated on the PS side to schedule diverse users, preserve privacy, eliminate redundant update transfers, and improve learning performance. In both entropy-based and LSE-based methods, we select diverse users to achieve a uniform data distribution, closely approximating full participation updates and minimizing error in (17).

VI. Conclusions

In this work, we study user scheduling and user representation estimation for OTA FL with EH devices over wireless fading MAC with highly heterogeneous data distribution. For both known and unknown data distribution scenarios, our framework proposes user scheduling based on the relations of users. We analyze the convergence rate for the OTA FL with EH devices and demonstrate the effect of user scheduling. Our entropy-based approach uses label distribution to reduce convergence error, while least squares estimation helps the PS infer user relationships for scheduling without compromising privacy. The numerical results from various setups demonstrate that performing user scheduling and update estimation significantly enhances performance, especially in highly heterogeneous distributions. This work can be extended by adapting to different EH device behaviors, varying cluster numbers, and performing estimation periodically. Building on this, a potential future direction is to implement clustered federated learning for the user clusters derived from our estimation.

APPENDIX A

We define:

$$\boldsymbol{w}(t+1) \triangleq \boldsymbol{\theta}_{PS}(t) + \frac{1}{|\mathcal{S}(t)|} \sum_{m \in \mathcal{S}(t)} \Delta \boldsymbol{\theta}_m(t), \tag{21}$$

$$\boldsymbol{v}(t+1) \triangleq \boldsymbol{\theta}_{PS}(t) + \frac{1}{M} \sum_{m=1}^{M} \Delta \boldsymbol{\theta}_m(t). \tag{22}$$

From (6), we have $\boldsymbol{\theta}_{PS}(t+1) = \boldsymbol{\theta}_{PS}(t) + \Delta \hat{\boldsymbol{\theta}}_{PS}(t)$. Using this, we can derive

$$\begin{aligned} &\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad = \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{w}(t+1) + \boldsymbol{w}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad = \|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{w}(t+1)\|_2^2 + \|\boldsymbol{w}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\qquad + 2\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{w}(t+1), \boldsymbol{w}(t+1) - \boldsymbol{\theta}^* \rangle. \end{aligned} \tag{23}$$

To bound these terms, we employ the following lemmas.

Lemma 1. *For the first and third terms in (23), we have*

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{w}(t+1)\|_2^2\right] \leq \frac{\eta^2(t)\tau^2 G^2}{K} + \frac{\sigma_z^2 N}{\alpha_t^2 K |\mathcal{S}(t)| \sigma_h^2},$$

and

$$\mathbb{E}\big[\langle \boldsymbol{\theta}_{PS}(t+1) - \boldsymbol{w}(t+1), \boldsymbol{w}(t+1) - \boldsymbol{\theta}^* \rangle\big] = 0.$$

Proof. The proofs are similar to Lemmas 1 and 3 in [5]. □

For the second term in (23), we proceed as follows:

$$\begin{aligned} &\|\boldsymbol{w}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad = \|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1) + \boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\quad = \|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1)\|_2^2 + \|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2 \\ &\qquad + 2\langle \boldsymbol{w}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^* \rangle. \end{aligned} \tag{24}$$

Lemma 2. *For the second term in (24), we have*

$$\begin{aligned} &\mathbb{E}\left[\|\boldsymbol{v}(t+1) - \boldsymbol{\theta}^*\|_2^2\right] \\ &\quad \leq (1 - \mu\eta(t)\ \ (\tau - \eta(t)(\tau-1)))\, \mathbb{E}\left[\|\boldsymbol{\theta}_{PS}(t)\ \ - \boldsymbol{\theta}^*\|_2^2\right] \\ &\qquad + (1 + \mu(1-\eta(t)))\, \eta^2(t) G^2 \frac{\tau(\tau-1)(2\tau-1)}{6} \\ &\qquad + \eta^2(t)(\tau^2 + \tau - 1)G^2 + 2\eta(t)(\tau-1)\Gamma. \end{aligned} \tag{25}$$

Proof. The proof follows the same argument in [5, Lemma 2]. □

Lemma 3. *For the first term in (24), we have*

$$\begin{aligned} &\mathbb{E}\left[\|\boldsymbol{w}(t+1) - \boldsymbol{v}(t+1)\|_2^2\right] \\ &\qquad = \left(\eta^2(t)\tau(\tau-1)LG + \eta(t)\tau\epsilon\right)^2. \end{aligned}$$

Proof. The proof is similar to [15, Lemma 1]. □

Lemma 4. *The third term in (24) is bounded by*

$$\begin{aligned} &\mathbb{E}\left[2\langle \boldsymbol{w}(t+1) - \boldsymbol{v}(t+1), \boldsymbol{v}(t+1) - \boldsymbol{\theta}^* \rangle\right] \\ &\qquad \leq \left(\eta^2(t)\tau(\tau-1)LG + \eta(t)\tau\epsilon\right) c, \end{aligned} \tag{26}$$

for some constant $c \geq 0$, which is related to Γ, G and μ.

Proof. The proof follows a similar line as Lemma 1 in [15]. □

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of the 20th Int. Conf. on Artificial Intelligence and Statistics*, vol. 54. PMLR, Apr 2017, pp. 1273–1282.

[2] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1622–1658, Third Quarter 2021.

[3] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.

[4] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[5] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. V. Poor, "Blind federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5129–5143, Aug. 2021.

[6] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, Mar. 2015.

[7] C. Chen, Y.-H. Chiang, H. Lin, J. C. Lui, and Y. Ji, "Joint client selection and receive beamforming for over-the-air federated learning with energy harvesting," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1127–1140, 2023.

[8] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Over-the-air federated learning with energy harvesting devices," in *IEEE Global Commun. Conf. (GLOBECOM)*, Rio de Janeiro, Brazil, Dec 2022, pp. 1942–1947.

[9] Q. An, Y. Zhou, Z. Wang, H. Shan, Y. Shi, and M. Bennis, "Online optimization for over-the-air federated learning with energy harvesting," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 7, pp. 7291–7306, July 2024.

[10] K. Zhang and X. Cao, "Federated learning with energy harvesting devices: An MDP framework," *arXiv preprint arXiv:2405.10513*, 2024.

[11] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[12] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[13] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *Proc. IEEE Int. Conf. Comput. Commun.*, Toronto, ON, Canada, July 2020, pp. 1698–1707.

[14] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi, "Clustered sampling: Low-variance and improved representativity for clients selection in federated learning," in *International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 3407–3416.

[15] R. Balakrishnan, T. Li, T. Zhou, N. Himayat, V. Smith, and J. Bilmes, "Diverse client selection for federated learning via submodular maximization," in *International Conference on Learning Representations (ICLR)*, Apr. 2022, pp. 1–18.

[16] S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey and implications," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 443–461, Third Quarter 2011.

[17] B. Tegin and T. M. Duman, "Federated learning with over-the-air aggregation over time-varying channels," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5671–5684, Aug. 2023.

[18] O. Aygün, M. Kazemi, D. Gündüz, and T. M. Duman, "Over-the-air federated edge learning with hierarchical clustering," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 17 856–17 871, Dec. 2024.

[19] A. Lutz, G. Steidl, K. Müller, and W. Samek, "Optimizing federated learning by entropy-based client selection," *arXiv preprint 2411.01240*, 2024.

[20] Y. LeCun, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*, 1998.

[21] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[22] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, University of Toronto, Department of Computer Science, Toronto, ON, Canada, 2009.

[23] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *International Conference on Learning Representations (ICLR)*, May 2021, p. 1–36.